Bioinformatics applied to gene transcription regulation

Gioia Altobelli

Department of Endocrinology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK

(Correspondence should be addressed to G. Altobelli; Email: g.altobelli@qmul.ac.uk)

Abstract

Understanding regulation of gene transcription is central to molecular biology as well as being of great interest in medicine. The molecular syntax of the concerted transcriptional activation/repression of gene networks in mammal cells, which shape the physiological response to the molecular signals, is often unknown or not completely understood. Combining genome-wide experiments with in silico approaches opens the way to a systematic comprehension of the molecular mechanisms of transcription regulation. Diverse bioinformatics tools have been developed to help unravel these mechanisms, by handling and processing data at different stages: from data collection and storage to the identification of molecular targets and from the detection of DNA motif signatures in the regulatory sequences of functionally related genes to the identification of relevant regulatory networks. Moreover, the large amount of genome-wide scale data recently produced has attracted professionals from diverse backgrounds to this cutting-edge realm of molecular biology. This mini-review is intended as an orientation for multidisciplinary professionals, introducing a streamlined workflow in gene transcription regulation with emphasis on sequence analysis. It provides an outlook on tools and methods, selected from a host of bioinformatics resources available today. It has been designed for the benefit of students, investigators, and professionals who seek a coherent yet quick introduction to in silico approaches to analyzing regulation of gene transcription in the post-genomic era.

Journal of Molecular Endocrinology (2012) 49, R51–R59

Introduction

Completion of the human genome reference and development of next-generation sequencing (NGS) has led to an exciting era in which the complexity of biomedical research can be tackled in a new way, requiring multidisciplinary expertise at multiple levels (Noble 2002, Quackenbush 2006b, Navlakha & Bar-Joseph 2011). Bioinformatics, as an integrative discipline in this field, increasingly occupies a larger portion of the biomedical stage. Formerly seen as a service in wet laboratories or as a specialization in computer science departments, it is presently an intrinsic component of scientific discovery (Noble 2002). Significant benefits can derive from both sharing experience in a multidisciplinary team and employing individuals who integrate interdisciplinary expertise. Thus, it is important to prepare a new generation of biologists to master computational and quantitative skills (Bialek & Botstein 2004, Pevzner 2004, Wingreen & Botstein 2006, Dudley & Butte 2009, Pevzner & Shamir 2009) as well as knowledge in specialized areas of biomedicine. Moreover, creating and maintaining appropriate infrastructures helps in unifying the heterogeneity of post-genomic data and dedicated bioinformatics toolkits. Large resources for data handling and manipulation, for example, open interactive systems such as Galaxy (Giardine et al. 2005), BioGPS (Wu et al. 2009), and the open source project R-Bioconductor (Gentleman et al. 2004), are highly welcome and point to a successful, cooperative model. However, most efforts cannot yet fully rely on such coherent supporting frameworks. Instead, expertise and knowledge of heterogeneous tools are required (Bialek & Botstein 2004, Pevzner 2004, Wasserman & Sandelin 2004, Quackenbush 2006b, Wingreen & Botstein 2006, Dudley & Butte 2009, Pevzner & Shamir 2009). Here, I want to collect a few reference landmarks for technology professionals who want to enter the post-genomic biology era and for biomedical investigators and students who want to build bioinformatics expertise on their own. Given the vastness of the subject, I won’t be able to cover all areas of interest, nor provide deep insight into the complex issues underlying tool conception and development. By focusing on a streamlined analysis workflow, I want to stimulate and support the reader’s intellectual curiosity and desire to develop individualized skills.
Background

The physiological responses of cells to any molecular signal are shaped by the molecular syntax of a concerted transcriptional activation/repression of many genes, referred to as regulatory networks. The major molecular events that underlie gene transcription primarily include the recognition of specific DNA elements known as transcription factor binding sites (TFBSs) by transcription factor (TF) proteins, as well as the recognition of cofactors that specifically link these protein–DNA binding events to the transcriptional molecular machinery (Wasserman & Sandelin 2004, Juven-Gershon et al. 2008). Besides, micro-RNA circuitry and a variety of posttranslation modifications, including the topical epigenetic markers, further modulate gene transcription with either functional or dysfunctional effect (Bartel 2009, Fabbri & Calin 2010, Portela & Esteller 2010, Lister et al. 2011). The packaging of DNA sequences around histone proteins to form nucleosomes and chromatin fibers (Hayes & Hansen 2001), that is, the structural organization of chromatin in the cell nuclei, masks/reveals TFBSs under different physiological conditions (Li et al. 2007, Park 2009). Various epigenetic modifications further alter the chromatin conformation, likely combining with each other to generate chromatin states (Ernst & Kellis 2010). In addition to a first layer of signals constituted by TFBS motifs, we can therefore picture a second layer of signals, defined by CTCF-binding sites that indicate active/inactive DNA sequence domains, or chromatin boundaries (Jothi et al. 2008, Jeziorska et al. 2009, Phillips & Corces 2009, Botta et al. 2010, Tolstorukov et al. 2011).

The architecture of a typical gene promoter is not rigorously defined, although a functional regulatory sequence comprises core elements (core promoter) and several enhancer/silencer elements scattered at various distances from the annotated transcriptional start sites (Wasserman & Sandelin 2004, Carninci et al. 2005, Halfon 2006, Juven-Gershon et al. 2008, Dollfini et al. 2009), as shown in Fig. 1. Enhancers often lie several kilo bases from the core, as observed in regulation by nuclear receptors (Lin et al. 2007, Kuttippurathu et al. 2011, Navlakha & Bar-Joseph 2011), and may lie in intragenic regions (mainly introns) as well as in the upstream sequence of a gene. The ‘linear’ architecture of the promoter (Fig. 1) in fact reflects the tri-dimensional structure of chromatin packaging, the conformation of which is consistent with a dynamic fractal globule able to fold/unfold any genomic locus, as recently revealed by Hi-C technique and Monte Carlo simulations (Lieberman-Aiden et al. 2009). The organization of chromatin on a genome scale deeply affects gene transcription regulation (Li et al. 2007, Ernst et al. 2011, Tolstorukov et al. 2011, Delest et al. 2012). Further advancements in our understanding of gene transcriptional regulation will derive from the exploitation of structural findings on chromosome territories (Fraser & Bickmore 2007, Lieberman-Aiden et al. 2009, Shaw 2010), TF–TFBS binding events (Honig & Rohs 2011, Nikolova et al. 2011, Delest et al. 2012), and from a systemic approach (He et al. 2009, Ashworth et al. 2011).

Figure 1 Schematic view of a hypothetical eukaryotic gene regulatory region. Right hand arrows stand for the transcriptional start site (TSS). Geometric shapes represent various DNA elements or motifs: blue shapes, various transcription factor binding sites; red trapezoid, the proximal core. Black spots represent chromatin boundaries such as CTCF-binding motifs. In higher dimensions (2D, top), the distal elements (1D, bottom far left) may be located closer to the core promoter and are therefore able to physically interact with the general transcriptional machinery. Multiple transcripts of a gene may be found in databases. Multiple TSSs have been observed for specific genes involved in developmental functions. Depending on the repository/species, the TSS may be associated with just one canonical transcript.
ChIP-seq technology has enabled the unbiased identification of all DNA sequences (TFBS) bound by a specific protein of interest (TF) (Park 2009, Liu et al. 2010) and is also successfully used in epigenomics (Bock & Lengauer 2008, Park 2008, Huss 2010, Lim et al. 2010, Ongenaert 2010). In silico approaches are currently applied in combination with genome-wide experimental techniques (D’Haeseleer 2006a, Elnitski et al. 2006, Dufour et al. 2007, Bickel et al. 2009, Gazdag et al. 2009, Delacroix et al. 2010).

The methods developed for detection of DNA motifs in the regulatory sequences of genes were devised in the framework of information theory (Hertz et al. 1990, Hertz & Stormo 1999), at a time when no structural clue about chromatin packaging was available. In this theory, the DNA molecule is simply a long word composed of four letters, the A G T C nucleotides, where the motif is hidden (Hertz et al. 1990, Hertz & Stormo 1999, Stormo 2000, Wasserman & Sandelin 2004, D’Haeseleer 2006a,b). The characterization of the (non-coding) DNA regions in terms of functional CRMs remains a computational challenge, especially for compositional signals, involving the combinatorial activation of different species of TFBSs (Bluthgen et al. 2005, Pierstorff et al. 2006, Van Loo & Marynen 2009).

A comprehensive probabilistic framework for DNA motifs discovery, devised by van Nimwegen (2007), provides a unifying view of the sequence analysis approach to gene transcription regulation. The in silico determination of chromatin boundaries has been attempted from experimental chromatin signatures (Won et al. 2008) and nucleosome positioning prediction (Ioshikhes et al. 2006, Segal et al. 2006). The first systematic characterizations of chromatin states based on the in silico integration of a variety of epigenetic marks in human T-cells and Drosophila have been recently provided by Ernst & Kellis (2010), Ernst et al. (2011) and Riddle et al. (2011).

The presence of repeated, similar signals in the same non-coding DNA region, that is, the statistical overrepresentation of a DNA motif, especially when the region and/or the signal are conserved across the species, is key to inferring a biological role for this DNA element. This is a key criterion for detecting functional motifs, referred to as phylogenetic footprinting (Wasserman & Sandelin 2004, D’Haeseleer 2006a,b, Hannenhalli 2008) and implemented in several dedicated bioinformatics tools (see below). Moreover, statistical overrepresentation of (combinations of) DNA motifs in many of the regulatory regions of a set of co-expressed genes may indicate that these gene targets are co-regulated. In this set of regulatory regions, specific combinations of DNA motifs, henceforth ‘DNA signatures’, will be detected with good statistics compared with backgrounds (see Fig. 2).

Biochemical pathways target different DNA sequence elements and implement the gene expression control that is key to the regulatory events. In principle, knowledge of the distribution of the functional DNA elements in the regulatory regions of a set of responsive genes should enable the inference of the relevant regulatory networks. Inference of these networks from expression microarray data can be attempted by reverse engineering (He et al. 2009, Ashworth et al. 2011). Comprehensive coverage of reverse engineering and systemic approaches to gene transcription (regulatory) networks can be found in He et al. (2009), Koyuturk (2010) and Ashworth et al. (2011), and references therein. Much more difficult, and as yet unproven, is a forward approach, i.e. inferring the regulatory networks from the distribution of (active) DNA motifs in the gene regulatory regions (Altabelli 2007). In general, reconstruction of gene networks is bound to provide a static picture of the actual biological pathways. Time evolution must be addressed with an appropriate experimental design.

Bioinformatics toolkits

We distinguish here two levels of data analysis: 1) the treatment of raw data, which leads to the identification of target genes, and the unraveling of the biological meaning (ontological or functional analysis) of these targets; and 2) the integration of different types of information into a unifying view, the reconstruction of gene networks and biological pathways, and the
generation of regulative hypotheses. Sequence analysis may be performed at level 1, helping in identifying TFBs from ChIP-Seq data, for example, and/or at level 2 as in the case study below. In the following, we illustrate analysis tools for a hypothetical workflow that includes whole-genome techniques (expression microarray and ChIP-seq) and provide a concise applicative example of sequence analysis of regulatory regions of these gene targets. Even though many computational tools are available, the assumptions under which they have been developed may not fully apply to one’s specific purpose. Thus, the analytical strategy is in best practice tailored for the specific biological system and research goals, possibly by controlling experiment design and always by evaluating the strengths and limitations of the method used. Both microarray and NGS data can be analyzed using open software packages within the R-Bioconductor framework (Gentleman et al. 2004) and/or Galaxy (Giardine et al. 2005). Methods and strategies for ChIP-seq analysis can be found in Pepke et al. (2009). Zhang et al. (2011) provide an extensive review on the impact of NGS technologies in genomics and Marra et al. in functional genomics and epigenomics (Morozova & Marra 2008, Hirst & Marra 2010). For foundations of gene expression analysis, see for example Quackenbush (2001, 2002, 2003 and 2006a) and Hsiao et al. (2005).

Once a list of (possibly) co-regulated genes is available, the detection of DNA signatures (Fig. 2) may be attempted using more than one method from two classes: 1) the algorithms that search for known motifs by screening the sequence against collections of TFBs models and 2) the algorithms that discover DNA motifs by unraveling the statistical property of the sequence, referred to as de novo. TFBs are usually modeled by either consensus (Xie et al. 2005, D’Haeseleer 2006b) or matrix (Wasserman & Sandelin 2004, D’Haeseleer 2006a,b). The model-based methods are somewhat biased by the quality and type of the model used, yet provide a direct estimate of which TF is involved in regulation (Wasserman & Sandelin 2004). The de novo methods may spot novel motifs, yet require output interpretation; and both classes demand an appropriate definition of background in order to assign statistical significance (usually P-values) to their outputs. A background set of sequences is usually constituted by the entire set of promoters of all genes for the species of interest or, for example, by the regulatory sequences of the genes that are functionally opposite, i.e. downregulated vs upregulated, depending on the problem under investigation.

An excellent web toolkit for (proximal) regulatory region analysis is the matrix-based Pscan (Zambelli et al. 2009). It is both user friendly and statistically robust, enabling fast detection of DNA signatures for several different promoter sizes and for several species. Moreover, the source code can be downloaded for stand-alone usage. Tools for de novo motif discovery, that is from the second class of algorithms mentioned earlier, are often used in analysis of ChIP-seq experiments: WEEDER (Pavesi et al. 2004, Tompa et al. 2005), MEME (Bailey & Elkan 1994), MDScan (Liu et al. 2002), and W-ChIPMotifs (Jin et al. 2009) are equipped with web interfaces. Convenient integrative frames for launching multiple DNA discovery algorithms (two or more of the above-mentioned) are TAMO (Gordon et al. 2005), by shell scripting, and the web-based CompleteMOTIFs (Kuttippurathu et al. 2011). Motif discovery algorithms eventually need to be coupled to other tools, either to link the discovered motifs to known TFBs or to classify them using, for example, clustering algorithms (Bickel et al. 2009, Hackenberg et al. 2011). STAMP (Mahony & Benos 2007), a user-friendly tool that may serve this purpose, is also handled by the latest version of TAMO (Gordon et al. 2005). The more recent Tmod tool kit (Sun et al. 2010) allows the integration of 12 different motif discovery algorithms for Windows operating systems. RSAT (Thomas-Chollier et al. 2008) is a comprehensive interactive platform that supports hundreds of genomes and implements a variety of approaches to regulatory sequence analysis including unique visualization features. Based on multispecies alignments, OPOSSUM (Ho Sui et al. 2005) and CORG (Dieterich et al. 2005) are the major resources for investigating distal motifs, providing both collections of conserved regulatory regions and conserved motifs. The method implemented in REDUCE Suite v2 algorithm (Foat et al. 2006) instead models the interaction between TF and TFBs, enabling the inference of the TF sequence-specific binding affinity from a single expression experiment.

Gene annotation enrichment tools help in identifying the relevant (altered) biological processes in genome-wide experiments. The 68 gene enrichment tools developed for the functional analysis of gene expression data are compared by Huang da et al. (2009a,b). Enrichment tools heavily rely on the quality of annotation databases, which are incomplete and biased by construction (Huang da et al. 2009a, Liberzon et al. 2011). Comprehensive, web-based platforms are DAVID (Huang da et al. 2009b), GSEA (Subramanian et al. 2005, Liberzon et al. 2011), also available as an R-Bioconductor package and as a desktop application, and GREAT (McLean et al. 2010), which has been developed for unbiased functional annotation of genomic regions in human and mouse genomes. The latter is therefore recommended for annotating ChIP-seq-derived targets. For an extensive catalog of bioinformatics resources and web tools, the reader is referred to the dedicated annual publication by Brazas et al. (2011).
In order to tackle the complexity of estrogenic regulation in breast cancer cells \textit{in silico}, a large sequence analysis workflow was set up as in Fig. 3 (Altobelli 2007). Lists of estrogen target genes were collected and organized in a manually curated, in-house database, by selecting those that exhibited an early response to estradiol in breast cancer cells (up/downregulation within 4 h), from gene expression microarray and gene-by-gene experiments published up to year 2005. The workflow combined approaches for DNA sequence analysis of proximal regions ($\leq$ 1 kbp) with a method that enabled the investigation of distal (up to 15 kbp) conserved nucleotide blocks. The input, which comprises several lists of estrogen-responsive genes and their relevant regulatory regions, was taken from a manually curated, in-house database (based on published experiments up to year 2005). The multiple outputs of this workflow were 1) sub-lists of genes associated by the presence of (at least) one motif/TBFS, 2) overrepresented motifs and their genomic localizations, 3) lists of putative transcription factors, and, after data integration, 4) regulatory hypotheses. Among other DNA signatures, several conserved GATA family motifs were detected in multiple occurrences in the upstream region of a small group of developmental genes including cyclin-G2 (CCNG2). Two (novel) degenerated ERE motifs were also detected in the proximity of two conserved GATA3 binding sites in the CCNG2 upstream regulatory region (Altobelli 2007). Cyclin-G2, a cell cycle inhibitor, is strongly downregulated in breast cancer cells and is a primary ER target gene (Stossi et al. 2006). The co-localized ERE and GATA motifs suggest that an interaction between the ER and GATA3 may be implicated in the regulation of the above-mentioned group of developmental genes and that this might play a role in mammary gland development and cancer.
differentiation (Chou et al. 2010). This interaction might, in fact, contribute to the shaping of the lactiferous ducts, determining the fate of luminal cells, which express estrogen receptor (ER+), compared to myo-epithelial cells, which do not express estrogen receptor (ER−). This may also have implications in the characterization of types of breast cancer (ER+ vs ER− luminal-like types).

Genome-wide experiments in estrogen-responsive cells have recently enabled the detection of presumably direct targets of ERs, as well as a catalog of EREs and binding sites for putative cognate TFs (Carroll et al. 2005, Carroll & Brown 2006, Lin et al. 2007). This makes it possible, in principle, to validate all in silico outcomes and to refine the workflow accordingly.

Declaration of interest
The author declares that there is no conflict of interest that could be perceived as prejudicing the impartiality of the review.

Funding
This research did not receive any specific grant from any funding agency in the public, commercial, or not-for-profit sector.

Acknowledgements
The author thanks all the students and colleagues whose persistent questions eventually prompted to write this work as well as Prof. R. Bargi, Prof. R. Flower, and Mr. J. Maflin for kindly proofreading the manuscript.

References


Evans RM, Blanchette M & Giguere V 2007 Genome-wide regulation of transcription by the orphan nuclear receptors ERRZ and γ. Cell Metabolism 5: 345–356. (doi:10.1016/j.cmet.2007.03.007)

Evans RM, Blanchette M & Giguere V 2007 Genome-wide orchestration of cardiac functions by the orphan nuclear receptors ERRα and γ. Cell Metabolism 5: 345–356. (doi:10.1016/j.cmet.2007.03.007)


Received in final form 15 June 2012
Accepted 21 June 2012
Made available online as an Accepted Preprint 21 June 2012